Chapter 6

Recursive Partitioning and Tree-Based Methods

6.1 Introduction

Recursive partitioning is the process for constructing a *decision tree*, where for each node we decide to split into two child nodes, or not to split. It is the key to the nonparametric statistical method of *classification and regression trees* (CART) introduced by Breiman, Friedman, Olshen and Stone, 1984.

The algorithm asks a series of hierarchical *Boolean* questions. For a continuous variable X_j, whether or not $X_j \leq \theta_i$ for some threshold value θ_i. For a categorical variable X_k with state space $\{\theta_1, \ldots, \theta_K\}$, whether or not $X_k \in S$, where S is a strict subset of $\{\theta_1, \ldots, \theta_K\}$.

Let Y be the variable to be predicted and $X_1, \ldots, X_r$ the collection of predictors. The output (Y) is a *class* variable; $Y \in \mathcal{C} = \{C_1, \ldots, C_L\}$. If $X_1, \ldots, X_r$ are continous variables, then the input space is $\mathbb{R}^r$ and, following the answers to successive questions, the input space is partitioned into a number of non-overlapping hyper-rectangles. To each hyper-rectangle is associated a class from $\mathcal{C}$, which may be the maximum likelihood estimator of Y based on the answers to the questions.

6.2 Classification Trees

A *classification tree* is the result of asking an ordered sequence of questions, where the next question in the sequence depends on the answers to the previous questions of the sequence. The sequence terminates in a prediction of the class.

The starting point is the *root node* and consists of the entire learning set $\mathcal{L}$. A node is a subset of the variables, which can be *terminal* or *non-terminal*. A *non-terminal* node is a node which splits into two child nodes. The binary split is determined by a Boolean condition on the value of a single variable, where the condition is either satisfied ("yes") or not satisfies ("no") by the observed value of the variable. A *terminal* node is a node that does not split.

All observations in $\mathcal{L}$ that have reached a particular (parent) node and satisfy the condition drop down to one of the two child nodes; the remaining observations drop down to the other child node.

In this way, each observation in $\mathcal{L}$ drops down to one of the terminal nodes.

There may be more than one terminal node with the same class label. A single-split tree with only two terminal nodes is called a *stump*. The set of all terminal nodes is a *partition* of the data; each datum will belong to exactly one terminal node.

Example Suppose we have two input variables X_1 and X_2.

- Q(root): $X_2 \leq \theta_1$? yes/no
- Q(yes): $X_1 \leq \theta_2$? yes/no
- Q(no): $X_2 \leq \theta_3$? yes/no
- Q(no)(yes): $X_1 \leq \theta_4$? yes/no

DRAW PICTURE OF THE TREE - IT HAS 5 TERMINAL NODES.

The space is split into 5 regions: Assume $(X_1, X_2) \in [0,1]^2$ and $\theta_i \in [0,1]$ for $i = 1,2,3,4$, then the 5 rectangles are $R_1 = [0,\theta_2] \times [0,\theta_1]$, $R_2 = [\theta_1,1] \times [0,\theta_1]$, $R_3 = [0,\theta_4] \times [\theta_1,\theta_3]$, $R_4 = [\theta_4,1] \times [\theta_1,\theta_2]$, $R_5 = [0,1] \times [\theta_3,1]$.

DRAW A PICTURE OF $[0,1] \times [0,1]$ PARTITIONED INTO RECTANGLES.

It is clear that categorical variables and ordinal variables can also be included; ordinal variables (which take values in a set $1,\ldots,N$ which represent an ordering are included in exactly the same way; the questions are of the form $X \leq \theta$ for some value of θ. For categorical variables, if a variable has M distinct categories represented in the data at the node, labelled $l_1,\ldots,l_M$, the set $\mathcal{S}$ of splits is simply the number of ways of partitioning into two non-empty subsets. There are $2^{M-1}-1$ ways of doing this. For example, if $M = 4$, there are $2^3 - 1 = 7$ possible splits: $(\{l_1\},\{l_2,l_3,l_4\})$, $(\{l_2\},\{l_1,l_3,l_4\})$, $(\{l_3\},\{l_1,l_2,l_4\})$, $(\{l_4\},\{l_1,l_2,l_3\})$, $(\{l_1,l_2\},\{l_3,l_4\})$, $(\{l_1,l_3\},\{l_2,l_4\})$, $(\{l_1,l_4\},\{l_2,l_3\})$.

Cleveland Heart Disease Data The data file `cleveland.data` from the UCI repository

`www.ics.uci.edu/mlearn/databases/heart-disease`

contains data obtained from a heart disease study conducted by the Cleveland Clinic Foundation. The response variable is `diag` (diagnosis of heart disease: `buff` = healthy, `sick` = heart disease). There were 303 patients in the study. There are 13 input variables: `age` (in years), `gender` (male / female), `cp` (chest pain type: `angina` = typical angina, `abnang` = atypical angina, `notang` = non-anginal pain, `asympt` = asymptomatic), `trestbps` (resting blood pressure), etc .....

Choosing the Quesions Each question splits the population of the node in two. When we are learning a classification tree (i.e. a list of questions), we choose the question which gives the greatest Kullback Leibler information.

So, if we have two classes, $\mathcal{C} = \{C_0, C_1\}$, where the class index is the value of Y, p the proportion for which $Y = 1$ and $1 - p$ the proportion for which $Y = 0$, let p_{11} be the proportion of those who answer 'yes' and $Y = 1$, p_{10} those who answer 'yes' and $Y = 0$, and p_{01} those who answer 'no' and $Y = 1$, p_{00} those who answer 'no' and $Y = 0$. The Shannon Information Gain is:

$$\sum_{i=0}^{1}\sum_{j=0}^{1} p_{ij} \log \frac{p_{ij}}{p_{i+}p_{+j}}$$

where $p_{i+} = p_{i0} + p_{i1}$ and $p_{+j} = p_{0j} + p_{1j}$. The question which gives the greatest Shannon Information Gain for each node is asked, until no question will give an appreciable increase in SIG. There are other possible criteria for choosing the question; SIG has good properties, which we'll discuss next.

6.3 Shannon Entropy and Information

We now to show how the negative of *Shannon entropy* gives a convincing approach to the amount of information given by the answer to a question if we know the probability distribution and why, when assessing the amount of information gained, the *Kullback-Leibler* divergence is a useful quantity.

In the following, we consider the *parameter space* $\Theta = \mathcal{C} = \{C_1, \ldots, C_L\}$, the set of possible classes.

Definition 6.1 (Shannon Entropy)**.** *For a distribution with density π over a parameter space Θ, the negative of the* Shannon entropy *is defined as:*

$$\mathcal{E}(\pi) := -\int_{\Theta} \pi(\theta) \log \pi(\theta) d\theta.$$

We follow Lindley by taking the negative of this quantity, which we call the *information* in the distribution:

$$\mathcal{I}(\pi) = -\mathcal{E}(\pi) = \int_{\Theta} \pi(\theta) \log \pi(\theta) d\theta.$$

The negative sign in Shannon's definition is due to the fact that he is considering the *opposite* of information; Shannon's entropy is a measure of *disorder*.

Shannon gives reasons why this is a good measure and we follow Lindley's description of Shannon's motivational arguments.

In the discussion here, Θ is finite and π_Θ is a probability mass function. In the absence of any prior information about classes, we can take $\pi(\theta) = \frac{1}{L}$ (uniform distribution) for each $\theta \in \{1, \ldots, L\}$ (the

class labels). A priori, $\mathcal{I}(p) = \sum_\Theta \pi(\theta) \log \pi(\theta))$, then the amount of information, say I, can be measured by the amount of additional information required before the value of θ is known.

This information could be provided in two stages:

Stage 1 Let $\Theta_1 \subset \Theta$ be a non-empty, strict subset of Θ where $\sum_{\theta \in \Theta_1} \pi_\Theta(\theta) \neq 0$ or 1. Suppose the experimenter is told whether $\theta \in \Theta_1$ or $\theta \in \Theta \backslash \Theta_1$. The prior distribution over $(\Theta_1, \Theta \backslash \Theta_1)$ is $(\Pi, 1-\Pi)$, where $\Pi = \sum_{\theta \in \Theta_1} \pi_\Theta(\theta)$.

In the second stage, suppose the experimenter is told the value of θ; the information provided is I_2 if $\theta \in \Theta_1$, or I_3 if $\theta \in \Theta \backslash \Theta_1$. The distributions over Θ_1 and $\Theta \backslash \Theta_1$ are $\frac{\pi_\Theta(\theta)}{\Pi}$ and $\frac{\pi_\Theta(\theta)}{1-\Pi}$ respectively.

Shannon requires that the information provided in the first stage and that the average amount in the second stage add up to the total information; that is:

$$I = I_1 + \Pi I_2 + (1 - \Pi) I_3.$$

This requirement is the fundamental postulate of Shannon.

Shannon proves that (apart from arbitrary multiplicative constant) $I(\pi) = \sum_{\theta \in \Theta} \pi_\Theta(\theta) \log \pi_\Theta(\theta)$ is the *only* function satisfying this property (together with a mild continuity property).

We can see that I, thus defined, has this property;

$$\begin{aligned}
I_1 &= \Pi \log \Pi + (1 - \Pi) \log(1 - \Pi) \\
I_2 &= \sum_{\theta \in \Theta_1} \tfrac{\pi_\Theta(\theta)}{\Pi} \log \tfrac{\pi_\Theta(\theta)}{\Pi} = \tfrac{1}{\Pi} \left(\sum_{\theta \in \Theta_1} (\log \pi(\theta) - \log \Pi \right) \\
I_3 &= \sum_{\theta \in \Theta \backslash \Theta_1} \tfrac{\pi_\Theta(\theta)}{1-\Pi} \log \tfrac{\pi_\Theta(\theta)}{1-\Pi} = \tfrac{1}{1-\Pi} \sum_{\Theta \backslash \Theta_1} \pi_\Theta(\theta) \left(\log \pi_\Theta(\theta) - \log(1 - \Pi) \right)
\end{aligned}$$

and the identity $I = I_1 + \Pi I_2 + (1 - \Pi) I_3$ follows directly. Shannon also shows that this is the *only* function of π_Θ for which this is satisfied for arbitrary π_Θ and $\Theta_1 \subset \Theta$.

After the experiment has been performed, a result x observed and the distribution over Θ updated to $\pi_{\Theta|X}(.|x)$, the information is:

$$\mathcal{I}(\pi_{\Theta|X}(.|x)) = \int_\Theta \pi_{\Theta|X}(\theta|x) \log \pi_{\Theta|X}(\theta|x) d\theta$$

and the information *gain* is:

$$\mathcal{K}(x) = \mathcal{I}(\pi_{\Theta|X}(.|x)) - \mathcal{I}(\pi_\Theta).$$

We assume that, given a true parameter value θ, the outcome x of an experiment is governed by a probability distribution $p_{X|\Theta}(.|\theta)$.

The information difference depends on the observation x. If we are choosing between different experiments (in this case questions to be asked), then clearly we do not know the outcome before we

carry out the experiment! We therefore average the information difference over all outcomes for an experiment to get a suitable measure:

$$\begin{aligned}\int \mathcal{K}(x)p_X(x)dy &= \int p_X(x)\int(\pi_{\Theta|X}(\theta|x)\log\pi_{\Theta|X}(\theta|x)-\pi_\Theta(\theta)\log\pi_\Theta(\theta))d\theta dx\\ &= \int\int\left(p_X(x)\frac{\pi_\Theta(\theta)p_{X|\theta}(x|\theta)}{p_X(x)}\log\frac{\pi_{\Theta|X}(\theta|x)p_X(x)}{p_X(x)}-\pi_\Theta(\theta)\log\pi_\Theta(\theta)\right)d\theta dx\\ &= \int\int\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)\log\frac{\pi_{\Theta|X}(\theta|x)p_X(x)}{p_Y(x)\pi_\Theta(\theta)}d\theta dx=\mathbb{D}_{KL}(\pi_{\Theta|X}p_X\|\pi_\Theta p_X). \quad (6.1)\end{aligned}$$

(Here $\pi_\Theta p_{X|\Theta}$ is the joint distribution over parameter space / state space).

This is the Kullback-Leibler divergence between the *joint* distribution $\pi_\Theta p_{X|\Theta}$ over $\Theta\times\mathcal{X}$ and the product distribution $\pi_\Theta p_X$ over $\Theta\times\mathcal{X}$ (if the parameter and observation were independent, the Kullback-Leibler divergence would be zero; the experiment would provide no information).

The Kullback-Leibler divergence has several important properties, which indicate that it is useful for measuring the gain of information from an experiment. Firstly, if f and g are two probability distributions over a state space $\mathcal{X}$, then $\mathbb{D}_{KL}(f\|g)\geq 0$, where the inequality is strict if f and g differ on a set of positive f probability. This follows from Jensen's inequality; if ϕ is a convex function and X a random variable with well defined expected value, then $\mathbb{E}[\phi(X)]\geq\phi(\mathbb{E}[X])$. The function $\phi(x)=-\log x$ is convex. Applying this to Kullback Leibler, this gives:

$$\begin{aligned}\mathbb{D}_{KL}(f\|g) &= \int f(x)\log\frac{f(x)}{g(x)}dx=-\int f(x)\log\frac{g(x)}{f(x)}dx\\ &\geq -\log\int f(x)\frac{g(x)}{f(x)}dx=-\log\int f(x)dx=-\log 1=0.\end{aligned}$$

Another property is the *additive* property, which was Shannon's basic reason for introducing the entropy functional. Let $\mathcal{E}$ denote an experiment which takes place in two parts, $\mathcal{E}=(\mathcal{E}_1,\mathcal{E}_2)$, where $\mathcal{E}_2$ is performed after $\mathcal{E}_1$. Let $\mathcal{K}_{\mathcal{E}_1}$ denote the average information provided by the whole experiment, $\mathcal{K}_{\mathcal{E}_1}$ the information provided by the first part and $\mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1}$ the additional information provided by the second, then

$$\mathcal{K}_{\mathcal{E}}=\mathcal{K}_{\mathcal{E}_1}+\mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1}.$$

This follows quite easily; $\mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1}$ is defined as the average information gain from the second part. Now, using $X=(X_1,X_2)$ to denote answers to two successive questions (or more generally two parts of an experiment) and $x=(x_1,x_2)$ to denote the two outcomes:

$$\begin{aligned}\mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1} &= \int_{\mathcal{X}_1} p_{X_1}(x_1) \int_{\mathcal{X}_2}\int_{\Theta} p_{X_2|\Theta,X_1}(x_2|\theta,x_1)\pi_{\Theta|X_1}(\theta|x_1)\log\frac{p_{X_2|\Theta,X_1}(x_2|\theta,x_1)\pi_{\Theta|X_1}(\theta|x_1)}{\pi_{\Theta|X_1}(\theta|x_1)p_{X_2|X_1}(x_2|x_1)}d\theta dx_2 dx_1 \\ &= \int_{\mathcal{X}}\int_{\Theta} p_{X|\Theta}(x|\theta)\pi_{\Theta}(\theta)\log\frac{p_{X|\Theta}(x|\theta)\pi_{\Theta}(\theta)p_{X_1}(x_1)}{p_{X_1|\Theta}(x_1|\theta)\pi_{\Theta}(\theta)p_X(x)}d\theta dx.\end{aligned}$$

The last line comes from taking $p_{X|\Theta} = p_{X_1,X_2|\Theta} = p_{X_2|X_1,\Theta}p_{X_1|\Theta}$ and $\pi_{\Theta|X_1} = \frac{\pi_\Theta p_{X_1|\Theta}}{p_{X_1}}$. From this:

$$\mathcal{K}_{\mathcal{E}_1}+\mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1} = \int_{\mathcal{X}}\int_{\Theta}\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)\left(\log\frac{\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)p_{X_1}(x_1)}{\pi_\Theta(\theta)p_{X_1|\theta}(x_1|\theta)p_X(x)}+\log\frac{\pi_\Theta(\theta)p_{X_1|\Theta}(x_1|\theta)}{\pi_\Theta(\theta)p_{X_1}(x_1)}\right)d\theta dx = \mathcal{K}_{\mathcal{E}}.$$

We now consider the concept of *independent experiments*; two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, whose outcomes are observations of random variables X_1 and X_2, where both distributions have the same parameter space Θ, are said to be *independent* if $p_{X_1,X_2|\Theta} = p_{X_1|\Theta}p_{X_2|\Theta}$. That is, for any given parameter value θ, X_1 and X_2 are conditionally independent *conditioned on the value of the parameter*. Suppose $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2)$ where $\mathcal{E}_1$ is performed first and $\mathcal{E}_2$ is then performed. Let $\mathcal{E}_2(x_1)$ indicate the experiment $\mathcal{E}_2$, given that $\mathcal{E}_1$ gave outcome x_1; independence means that $\mathcal{E}_2(x_1) = \mathcal{E}_2$, which does not depend on x_1.

If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, then $\mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1} \leq \mathcal{K}_{\mathcal{E}_2}$, with equality if and only if $X_1 \perp X_2$ (i.e. they are *marginally* independent; $p_{X_1,X_2} = p_{X_1}p_{X_2}$). This is seen by a simple computation:

$$\begin{aligned}\mathcal{K}_{\mathcal{E}_2} - \mathcal{K}_{\mathcal{E}_2|\mathcal{E}_1} &= \int_{\Theta}\int_{\mathcal{X}}\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)\left(\log\frac{\pi_\Theta(\theta)p_{X_2|\Theta}(x_2|\theta)}{\pi_\Theta(\theta)p_{X_2}(x_2)}-\log\frac{\pi_{\Theta|X_1}(\theta|x_1)p_{X_2|\Theta,X_1}(x_2|\theta,x_1)}{p_{X_2|X_1}(x_2|x_1)\pi_{\Theta|X_1}(\theta|x_1)}\right)dxd\theta \\ &= \int_{\Theta}\int_{\mathcal{X}}\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)\left(\log\frac{\pi_\Theta(\theta)p_{X_2|\Theta}(x_2|\theta)}{\pi_\Theta(\theta)p_{X_2}(x_2)}-\log\frac{\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)p_{X_1}(x_1)}{p_X(x)\pi_\Theta(\theta)p_{X_1|\Theta}(x_1|\theta)}\right)dxd\theta \\ &= \int_{\Theta}\int_{\mathcal{X}}\pi_\Theta(\theta)p_{X|\Theta}(x|\theta)\log\frac{p_{X_1|\Theta}(x_1|\theta)p_{X_2|\Theta}(x_2|\theta)}{p_{X|\Theta}(x|\theta)}\frac{p_X(x)}{p_{X_1}(x_1)p_{X_2}(x_2)}dxd\theta \\ &= \int_{\mathcal{X}} p_X(x)\log\frac{p_X(y)}{p_{X_1}(x_1)p_{X_2}(x_2)}dx \geq 0.\end{aligned}$$

The expression in the last line is a Kullback-Leibler divergence, which is 0 if and only if $p_{X_1,X_2} = p_{X_1}p_{X_2}$.

This tells us (among other things) that if Experiment 2 is an independent repeat of Experiment 1, then the repetition is less informative, on average, than the original experiment.

Indeed, if we consider $\mathcal{E}_1, \mathcal{E}_2, \ldots$ a sequence of independent identical experiments and $\mathcal{E}^{(n)} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$, let $\mathcal{K}_n := \mathcal{K}_{\mathcal{E}^{(n)}}$, then $\mathcal{K}_n$ is a *concave* increasing function of n.

6.3.1 Tree-Growing Procedure

Some basic questions have to be answered:

1. How do we choose the Boolean conditions for splitting at each node? The choice of SIG is motivated by the fact that the sum of information gains from a sequence of questions is the same as the information gain if the multiple question were posed. This is a versatile choice, but not the only one.

2. Choice of criterion for when to split a parent node into two child nodes or when to decide if it is a terminal node.

3. Assigning a class to a node.

Node Impurity Functions Ideally, we would like all elements of a terminal node to belong to the same class, but this is not to be expected. Impurity is a measure of the amount of mixing in terminal nodes. Suppose that Y takes values in $\{1, \ldots, K\}$ (there are K possible classes). For node τ, we define the *node impurity function* as:

$$i(\tau) = \phi(p(1|\tau), \ldots, p(K|\tau))$$

where $p(i|\tau)$ is the proportion of class i observations in node τ. This is an estimate of $\mathbb{P}(X \in \Pi_i|\tau)$, probability that the observation is in class i given that the questions thus far place the observation at node τ.

The Shannon Information Gain is obtained by using

$$i(\tau) = -\sum_{k=1}^{K} p(k|\tau) \log p(k|\tau)$$

There are other possibilities; for example,

$$i_G(\tau) = \sum_{k \neq k'} p(k|\tau)p(k'|\tau) = 1 - \sum_{k} (p(k|\tau))^2.$$

i_G is the so-called Gini index. If classification is binary, then the entropy is

$$i(\tau) = -p \log p - (1-p) \log(1-p)$$

and the Gini index is:

$$i_G(\tau) = 2p(1-p).$$

6.4 Assigning classes to nodes: Estimating the Misclassification Rate

Suppose we have reached a node τ. The misclassification rate is:

$$R(\tau) = 1 - \max_k p(k|\tau).$$

For two classes, this is:

$$R(\tau) = 1 - \max(p, 1-p) = \min(p, 1-p).$$

For a tree, the mis-classification is based on the *terminal* nodes. If $\widetilde{\mathcal{T}}$ denotes the set of terminal nodes, then the true misclassification rate for the tree $\mathcal{T}$ is:

$$R(\mathcal{T}) = \sum_{\tau \in \widetilde{\mathcal{T}}} R(\tau) P(\tau)$$

where $P(\tau)$ is the probability that an observation is placed in (terminal) node τ. We use estimates (based on the learning set where classifications are known) to estimate $P(\tau)$ and $R(\tau)$ for each terminal node τ.

6.5 Pruning the Tree

The tree is grown according to a greedy algorithm; for each node, choose the question which gives the greatest increase in score for that node. This can lead to a tree that is too large. For tree pruning, we use a regularisation approach, starting at the terminal nodes and removing them if they do not represent sufficient gain over the parents. For a node τ, which is terminal in the current tree, we consider:

$$R_\alpha(\tau) = R(\tau) + \alpha$$

where $R(\tau)$ denotes the estimated mis-specification. Then

$$R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha|\widetilde{\mathcal{T}}|.$$

The term $\alpha|\widetilde{\mathcal{T}}|$ is a penalty on the tree size. For each α, we choose the subtree $\mathcal{T}_\alpha$ which minimises $R_\alpha(\mathcal{T})$. This gives $\mathcal{T}(\alpha)$. The tree $\mathcal{T}(\alpha)$ is not necessarily unique.

The chosen value of α determines the tree size. Although $\alpha \in [0, +\infty)$, the number of possible sub-trees of $\mathcal{T}$ is finite. We can consider α_1 the lowest value of α such that $\mathcal{T}(\alpha) \neq \mathcal{T}$ and let $\mathcal{T}_1 = \mathcal{T}(\alpha)$, α_2 the next lowest yielding $\mathcal{T}_2 = \mathcal{T}(\alpha_2)$ and so on. This gives a finite sequence of trees $\mathcal{T} \supset \mathcal{T}_1 \supset \mathcal{T}_2 \subset \ldots$.

Suppose a node τ in an optimal tree $\mathcal{T}$ has two terminal child nodes τ_L and τ_R, then $R(\tau) \geq R(\tau_L) + R(\tau_R)$ (we're using R to denote the estimates used to generate the tree). Now let $\mathcal{T}_1, \mathcal{T}_2, \ldots$ denote the

trees obtained by reducing $\mathcal{T}$ as α is increased. Let (τ_1, τ_2) denote the terminal nodes of $\mathcal{T}$ which are not in $\mathcal{T}_1$ (in case of ambiguity, we take a specific sequence of trees) and let $\tau \in \mathcal{T}_1$ denote the terminal node in $\mathcal{T}_1$ which is a non-terminal node in $\mathcal{T}$. For a node τ in a tree $\mathcal{T}$, we denote by $\mathcal{T}_\tau$ the subtree with root τ, going down to the terminal nodes of $\mathcal{T}$.

As long as $R_\alpha(\tau) \geq R_\alpha(\mathcal{T}_\tau)$, the subtree $\mathcal{T}_\tau$ has lower cost than terminating the tree at τ and hence it is retained.

Therefore, when

$$\alpha < \frac{R(\tau) - R(\mathcal{T}_\tau)}{|\widetilde{T}_\tau| - 1}$$

we retain $\mathcal{T}_\tau$. We can set

$$g_1(\tau) = \frac{R(\tau) - R(\mathcal{T}_{1,\tau})}{|\widetilde{T}_{1,r}| - 1} \qquad \tau \notin \mathcal{T}(\alpha_1)$$

where $\mathcal{T}_{1,\tau} = \mathcal{T}_\tau$ and $g_1(\tau)$ gives the critical value for α; when $g_1(\tau) \geq \alpha_1$ for each τ, we do not prune the terminal nodes.

The *weakest link* node $\widetilde{\tau}_1$ is the node in $\mathcal{T}_1$ that satisfies

$$g(\widetilde{\tau}_1) = \min_{\tau \in \mathcal{T}_1} g(\tau).$$

As α increases, $\widetilde{\tau}_1$ is the first node for which $R_\alpha(\tau) = R_\alpha(\mathcal{T}_\tau)$, so $\alpha_2 = g_1(\widetilde{\tau}_1)$. Recursively,

$$g_3(\tau) = \frac{R(\tau) - R(\mathcal{T}_{2,\tau}}{|\widetilde{\mathcal{T}}_{2,\tau}| - 1} \qquad \tau \in \mathcal{T}(\alpha_2), \quad \tau \notin \widetilde{\mathcal{T}}(\alpha_2)$$

and so on.

6.5.1 Choosing the best pruned subtree

Choosing the subtree requires good estimates of the misclassification rate. There are two approaches: for large data sets, using an independent test set is straightforward and computationally efficient. For small data sets, cross validation is recommended. Randomly assign the data into two sets of equal size, the learning set and the test set. Construct the tree using the learning set; estimate the misclassification rate using the test set.

At each stage, dropping down a level, let the chance of misclassification be p^*. We can consider each observation dropped down as a Bernoulli trial, from which we can compute the estimate of misclassification, together with a standard error.

Cross Validation Divide the data into V sets of approximately equal size, call them $D_1, \ldots, D_V$. Create V learning sets $\mathcal{L}_v = D \backslash D_v$. Use $\mathcal{L}_v$ to learn the classification tree $\mathcal{T}^v$. Fix the value of the complexity parameter α and let $\mathcal{T}^v(\alpha)$ be the best pruned subtree of $\mathcal{T}^v$, $v = 1, \ldots, V$. Drop each

observation of the vth test set down the tree $\mathcal{T}^v(\alpha)$ and let n_{ij}^v denote the number of observations class j that are classified as being of class i from test set v. Then $n_{ij}(\alpha) = \sum_{v=1}^{V} n_{ij}^v(\alpha)$. Set

$$R^{CV/V}(\mathcal{T}(\alpha)) = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1;j\neq i}^{K} n_{ij}(\alpha)$$